Running head: REASONABLE DOUBT

In press: *Psychology, Public Policy & Law*

Instructions on Reasonable Doubt: Defining the Standard of Proof and the Juror's Task

Mandeep K. Dhami[1], Samantha Lundrigan[2], and Katrin Mueller-Johnson[3]

[1]Middlesex University, London

[2]Anglia Ruskin University, Cambridge

[3]University of Cambridge

Send correspondence to:
Mandeep K. Dhami, PhD
Professor of Decision Psychology
Department of Psychology
Middlesex University
The Burroughs, Hendon
London, NW4 4BT

E-mail: m.dhami@mdx.ac.uk

Acknowledgements

The research presented in this paper was funded by a grant to MKD and KMJ from the Arts and Humanities Research Council. We would like to thank NAPP Plc for assisting with data collection.

Instructions on Reasonable Doubt: Defining the Standard of Proof and the Juror's Task

'Reasonable doubt' (RD) is the standard of proof used in criminal trials. It specifies the degree of belief in (or probability of) guilt required for conviction. As a principle of due process that aims to minimize the number of false convictions, RD is a stringent threshold. It has been theorized to be numerically equivalent to a .90 (or 90%) level of certainty following Blackstone's ratio of it being preferable to acquit 10 guilty persons rather than convict one innocent person (see Blackstone, 1765; Laudan, 2003; McCauliff, 1982; Newman, 1993; United States v. Fatico, 1978).

Researchers using direct rating scales to quantify people's interpretations of RD have shown that although some judges and (mock) jurors do interpret the standard at around .90 or above (e.g., Martin & Schum, 1987; Montgomery, 1998; Zander, 2000), others have much lower interpretations (e.g., Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Wright & Hall, 2007). In addition, there is considerable *inter*-individual variability in interpretations, meaning that different people interpret RD differently (for a review see Hastie, 1993). Recent research also points to *intra*-individual variability in interpretations such that an individual may interpret the standard as anywhere from around .50 to over .90 (Dhami, 2008; Lundrigan, Dhami & Mueller-Johnson, 2013, 2014; Mueller-Johnson, Dhami & Lundrigan, 2014).

Several jurisdictions have attempted to improve people's comprehension of the standard and bring its interpretation closer to that intended, as well as reduce variability in interpretations. These efforts have largely involved the development and introduction of judicial instructions that allow the judge to define RD to jurors using standardized terminology. To date, a variety of instructions on RD have been developed (see Heffer, 2006; Hemmens, Scarborough & Del Carmen, 1997; Power, 1999). Hemmens et al. (1997) identified at least six different instructions on RD used in the US federal courts of appeal, and

many more used across different US States (e.g., ‘doubt based on reason’, ‘actual and substantial doubt’, ‘doubt that can be articulated’, and ‘moral certainty of guilt’).

In the present paper, we measure the effectiveness of two specific instructions on RD that have been used and challenged in the US. These are described in more detail below. We examine the effect of the two instructions on people’s quantitative interpretations of the standard, as well as the degree of inter- and intra-individual variability in interpretations. In addition, we attempt to explain the effectiveness of these instructions by examining the effect of their precise wording on interpretations of RD.

It is important to study the language of RD instructions because criminal convictions can be appealed on the basis that the standard was defined inappropriately (i.e., leading to a reduced threshold for conviction; see Power, 1999; e.g., Cage v. Louisiana, 1990; Sullivan v. Louisiana, 1993). In addition, appeals courts have sometimes refused to reverse a decision when certain language has been used in the context of other language, implying that any problems with some parts of an RD instruction can be cancelled out or corrected by other parts (e.g., Victor v. Nebraska, 1994; United States v. Emalfarb, 1973; United States v. Miller, 1996). Before describing the RD instructions of interest in the present study, we review relevant theory and past research on the topic.

Reasonable Doubt as a Fuzzy Concept

Theory of Linguistic Probabilities

According to some researchers, standards of proof such as RD are akin to linguistic probabilities such as ‘very likely’ that are used to understand and communicate uncertainty (see Clermont, 1987, 2013; Dhami, 2008; Schum, 1986). The numerical equivalent of these phrases would lie along the 0-1 probability scale. Indeed, RD is said to be equivalent to a .9 probability. Wallsten and Budescu’s (1995) theory of linguistic probabilities borrows from Zadeh’s (1965) theory of fuzzy sets in mathematics, and states that these phrases can be

represented as *fuzzy* subsets of the probability interval (see also Budescu & Wallsten, 1995). A phrase such as RD would thus be characterized by a peak and spread of probabilities called a membership function (MF).

Figure 1 provides an example of a MF for an individual juror. Here, the degree of membership of RD at probabilities 0 to .5 is 0 because the juror believes that RD is *not at all* described by these very low probability values. The degree of membership of RD at probabilities .51 to .89 are closer to 1 because the juror believes that to *some degree* RD can be described by these mid-range values. The degree of membership of RD at probability .90 is 1 because the juror believes that RD can be described *absolutely* by this probability value. Finally, the degree of membership of RD at probabilities .91 to 1 are again 0 because the juror believes that RD is *not at all* described by these higher values. According to Wallsten and Budescu (1995), the MF of RD for this juror would peak at .9 and have a spread of .4.

FIGURE 1 ABOUT HERE

Comparing MF peaks across people demonstrates the extent of inter-individual variability in interpretations of RD, whereas an examination of the spread of an individual's MF demonstrates intra-individual variability in interpretations. Past research on linguistic probabilities has found that people have broad interpretations for most phrases in their linguistic probability lexicons, and that interpretations of the same phrase differ across people (e.g., Budescu, Weinberg, & Wallsten, 1988; Dhami & Wallsten, 2005; Erev & Cohen, 1990; Karelitz & Budescu, 2004; Zwick & Wallsten, 1989). Interpretations may also be affected by the context in which a phrase is used (e.g., Harris & Corner, 2011; Smithson, Budescu, Broomell, & Por, 2012; Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990).

Therefore, one goal of RD instructions should be to encourage peak interpretations of the standard to be close to that intended (i.e., .9) for all individuals. This also reduces inter-

individual variability. Another goal should be to reduce the spread of interpretations along the probability scale for each person, thus reducing intra-individual variability.

Past Research on Reasonable Doubt Instructions

There is a growing body of research on the effectiveness of RD instructions (for a review see Horowitz, 1997; Stoffelmayr & Diamond, 2000). However, most of this research has used quantitative measures of interpretations of RD that yield only point estimates on the 0-1 probability scale. Here, several researchers have compared the effect of specific RD instructions with a condition where the standard is left undefined. Some of these studies show that interpretations of RD are around .90 under some definitions compared to when the standard is left undefined (e.g., Horowitz & Kirkpatrick, 1996; Kerr, Atkin, Stasser, Meek, Holt & Davis, 1976; see also Elwork, Sales & Alfini, 1982). For example, in an early study, Kerr et al. (1976) used a simulated rape trial and found that when RD was defined as 'any doubt', mock jurors had significantly higher interpretations of the standard compared to when RD was undefined or when it was defined as 'substantial doubt'.

However, other studies do not necessarily support this beneficial view of instructions (e.g., Dhami, 2008; Koch & Devine, 1999; Kramer & Koenig, 1990; Montgomery, 1998; Mueller-Johnson et al., 2013; Wright & Hall, 2007). For instance, Koch and Devine (1999) found that participants rendered more guilty verdicts under the 'firmly convinced' instruction used in the US federal courts than when the standard was undefined, implying that the instruction did not lead to higher interpretations of RD.

Other researchers have compared the relative effect of different RD instructions on interpretations of the standard (e.g., Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Kagehiro & Stanton, 1985; Kerr et al., 1976; Nagel, 1979; Mueller-Johnson et al., 2013; see also Horowitz, 1990). These studies have found that some instructions lead people to adopt a more stringent standard while other instructions lead them to adopt a more lax standard. For

example, in the context of a hypothetical murder case where the strength of the evidence was manipulated to be weak or strong, Horowitz and Kirkpatrick (1996) compared interpretations of RD under four different instructions used in the US (i.e., 'firmly convinced', 'moral certainty', 'does not waiver or vacillate', and 'real doubt') as well as when the standard was left undefined. Mock jurors grouped into six person juries provided their interpretations of RD at both the pre- and post-deliberation stages. It was found that at both stages, and for both levels of evidence strength, the most stringent interpretations of RD were under the 'firmly convinced' instruction. The most lax interpretations of RD varied across conditions (i.e. 'moral certainty' in the strong evidence/pre and post-deliberation conditions; 'waiver/vacillate' in the weak evidence/post-deliberation condition; and undefined in the weak evidence/pre-deliberation condition).

Where instructions on RD could be of value is in efforts at reducing the variability in interpretations of the standard. Mueller-Johnson et al. (2014) captured the MFs of mock jurors' interpretations of RD under the 'firmly convinced' instruction and the 'sure' instruction which is used in the UK (see also Dhami, 2008). It was found that there was greater inter-individual variability in interpretations of RD under the 'sure' than 'firmly convinced' instruction or when the standard was undefined. In relation to intra-individual variability, both instructions led to a similarly high level of variability compared to when the standard was left undefined. Thus, despite the introduction of RD instructions, courts may often be ineffective in communicating the standard to jurors.

Unfortunately, much of the above literature does not directly investigate reasons for the observed (in)effectiveness of RD instructions. One potential explanation for why instructions on RD may fail to achieve their objective refers to the psycholinguistic complexity of these instructions (e.g., Charrow & Charrow, 1979; Dumas, 2000; Elwork et al., 1982; Heffer, 2006; Solan, 2001; Power, 1999). For instance, the grammatical structure of

instructions can be poor, and the instructions may be quite lengthy. In addition, although RD is not a phrase that is commonly used in everyday language, terms or phrases used to define RD are often similarly unusual, and these may themselves be used in the context of further unusual language. For example, the Supreme Court of Canada in R. Lifchus (1997) states that RD is not 'an imaginary or frivolous doubt', nor is it based on 'sympathy or prejudice.' It is doubt based on 'reason and commonsense', and does not mean 'probably guilty' or 'absolute certainty.' Thus, efforts to understand how RD instructions affect jurors' interpretations of the standard should systematically examine the precise language used to define the standard.

Reasonable Doubt Defined as 'Proof-Willing' and 'Doubt-Hesitate'

In the present study, we examine the effects of the language that appears in two different RD instructions used in 15 US States (e.g., Hemmens et al., 1997). For brevity, we refer to these instructions as 'proof-willing' and 'doubt-hesitate'. The 'proof-willing' instruction can be traced back to 1887 when the Supreme Court upheld its use in Hopt v. State of Utah. It is currently endorsed in five US States. For example, Ohio uses the following 'proof-willing' instruction: "Proof beyond a reasonable doubt is proof of such character that an ordinary person would be willing to rely and act upon it in the most important of the person's own affairs." (Ohio Jury Instructions, 2008, CR 417).

The 'doubt-hesitate' instruction dates back to at least 1922 (Newman, 1993; Posey v. State of Alabama, 1922) and is currently used in 10 US States. Pennsylvania uses the following 'doubt-hesitate' instruction: "A reasonable doubt is a doubt that would cause a reasonably careful and sensible person to pause and hesitate before acting upon a matter of importance in his or her own affairs." (Pennsylvania Suggested Standard Criminal Jury Instructions, 2008, 7.01).

Both instructions have been the subject of many legal debates (e.g., see Fortunator, 1996; Laudan, 2003; Power, 1999). However, these debates have largely centered on whether

it is appropriate to equate jury verdicts in criminal trials with important decisions made in ordinary life. Although this analogy is used in both the 'proof-willing' and 'doubt-hesitate' instructions, courts have generally favored the 'doubt-hesitate' over the 'proof-willing' instruction (e.g., Holland v. US, 1954; Poulson v. State of Texas, 1999; US v. Mars, 1977; Wainwright v. US, 1999). It is therefore unclear on what basis the courts prefer one instruction over the other.

In addition, the courts have not explicitly taken into account the fact that the phrases 'willing to act' and 'hesitate to act' are confounded with the words 'proof' and 'doubt', respectively. In the 'proof-willing' instruction, the phase 'willing to act' appears in conjunction with the word 'proof', whereas in the 'doubt-hesitate' instruction, 'hesitate to act' is combined with 'doubt'. It is conceivable that these phrases and words could have independent effects on how RD is interpreted or that they have an interactive effect.

Indeed, the language of the 'proof-willing' and 'doubt-hesitate' instructions is of particular interest because beyond defining the standard of proof, it appears to guide jurors in the tasks of evidence evaluation and conviction in different ways. In terms of evidence evaluation, the words 'proof' and 'doubt' appear to ask jurors to consider evidence as providing proof of guilt versus signifying doubt in guilt, respectively. In terms of conviction, the phrases 'willing to act' and 'hesitate to act' appear to ask jurors to perform an action (i.e., to convict) or not to act (i.e., not to convict), respectively. Thus, the 'proof-willing' instruction appears to ask jurors to consider evidence as providing proof of guilt and to act to convict (see e.g., Van Gundy v. State of Ohio, 1992 and Cooper v. State of Alaska, 2008). By contrast, the 'doubt-hesitate' instruction appears to ask jurors to consider evidence as signifying doubt in guilt and not act (or hesitate) to convict (see e.g., Harris v. State of Idaho, 2001 and Laramore v. State of Idaho, 2007). It is conceivable that by using different

terminology and by defining the jurors' tasks of evidence evaluation and conviction differently, the two instructions also have differential effects on jurors' interpretations of RD.

The Present Study

To our knowledge, no past research has empirically studied the 'proof-willing' and 'doubt-hesitate' instructions, and the present study represents a first attempt to do so. The study had four aims. The first was to compare the effect of the 'proof-willing' and 'doubt-hesitate' instructions on people's quantitative interpretations of RD. The second aim was to compare the degree of *inter*-individual variability in interpretations of the standard under these two instructions. The third was to compare the degree of *intra*-individual variability in interpretations of RD under the two instructions. The final aim was to explain the effect of the two instructions by examining how their precise language influences interpretations of RD. In particular, we measured how the phrases 'willing to act' and 'hesitate to act' and the words 'proof' and 'doubt' affect interpretations of the standard in isolation/independently of each other and in conjunction with one another.

A lack of past research on the 'proof-willing' and 'doubt-hesitate' instructions, as well as on the specific language contained in these instructions (i.e., 'willing to act', 'hesitate to act', 'proof', and 'doubt') precluded a priori directional hypotheses as to their effects on interpretations of RD. However, past research on RD instructions generally (see e.g., Hastie, 1993; Dhami, 2008) and past research on linguistic probabilities (see e.g., Budescu & Wallsten, 1995; Dhami & Wallsten, 2005) as reviewed above, led us to expect inter- and intra-individual variability in interpretations of RD under both instructions.

Method

Participants

The study involved 200 members of the jury eligible public recruited from a large pharmaceutical company in the UK that has a multi-building site and employs over 700

people in a wide range of roles from packers through clerical staff to medics. They volunteered to participate in return for a payment of £10. The mean age was 35.87 (*SD* = 10.68). Fifty-two percent of the sample was female, and 94.0% described their ethnicity as White. Fifty-three percent had a university-level education, and 89.5% were employed (the unemployed were unpaid interns and volunteers). Eight percent of the sample reported having served on a jury before. On average, participants rated the likelihood of them attending jury service if they were called to do so, as being 76.4% (*SD* = 29.8).

Design and Stimuli

We studied the effects of the two existing instructions using a 2 x 2 between-subjects factorial design. This resulted in four RD instructions, two of which represent the existing instructions (i.e., 'proof-willing' and 'doubt-hesitate') and two of which help to de-confound the phrases 'willing to act' and 'hesitate to act' from the words 'proof' and 'doubt'. The four instructions were written as follows:

- "Reasonable doubt is *proof* that would make a reasonable person *willing* to act in their most important affairs of life." (Hereafter called 'proof-willing')
- "Reasonable doubt is *doubt* that would make a reasonable person *hesitate* to act in their most important affairs of life." ('Doubt-hesitate').
- "Reasonable doubt is *proof* that would make a reasonable person *hesitate* to act in their most important affairs of life." ('Proof-hesitate').
- "Reasonable doubt is *doubt* that would make a reasonable person *willing* to act in their most important affairs of life." ('Doubt-willing').

To a native English speaker, the latter two definitions will appear to be grammatically inappropriate. However, we could find no other way to test the independent and relative effect of the precise language of the instructions. If we had added more (and different) language to improve the grammar of each instruction (e.g., "Reasonable doubt is

that *lack of* proof which would make a reasonable person hesitate to act in their most important affairs of life" or "reasonable doubt is that *amount of* doubt by which a reasonable person *would nevertheless* be willing to act in their most important affairs of life") this would have introduced confounds into the experiment, making it difficult to isolate the effects of the language of interest from the additional words (e.g., 'lack of', 'amount of', and 'would nevertheless). On the other hand, if we added language similarly to each instruction (e.g., "Reasonable doubt is *a level* of proof that would make a reasonable person *still* hesitate to act in their most important affairs of life" or "reasonable doubt is a *level of* doubt that would make a reasonable person *still* willing to act in their most important affairs of life), it would be difficult to disentangle the effects of the additional words (e.g., 'still') from the language of interest.

Importantly, we are not interested in comparing across all four instructions, but rather across the two existing instructions, and primarily interested in isolating the effects of the specific language i.e., 'willing to act', 'hesitate to act', 'proof', and 'doubt'. At the beginning of the study, the second author (who collected the data) told participants to ask for help if they did not fully understand the instructions or the task, and at the end of the study, she asked participants how well they had understood what was asked of them. None of the participants in the two non-existent conditions had questions or reported difficulties. The results also suggest a pattern of responses consistent with the idea that participants were not confused and unsystematically variable in their responses.

Measures

Participants' interpretations of RD were measured using two methods. One was the direct rating method. This yields a point numerical estimate and has been commonly employed in past research (see Dhami, 2008; Hastie, 1993; Horowitz & Kirkpatrick, 1996; Lundrigan et al., 2013, 2014; Montgomery, 1998). Here, after reading the RD instruction,

participants were asked: “How much [proof/doubt] would make you [hesitate/willing] to find someone guilty of a crime?” Responses were provided on a 0-100% scale (with 5% intervals; 21-points).

Since the direct rating method does not capture intra-individual variability in interpretations, we also used the Membership Function (MF) method (see Dhami & Wallsten, 2005; Karelitz & Budescu, 2004), which is associated with Wallsten and Budescu’s (1995) theory of linguistic probabilities. The MF method was restricted to the two existing RD instructions (i.e., ‘proof-willing’ and ‘doubt-hesitate’). Dhami (2008) first introduced this method for studying standards of proof. In the present study, after reading the instruction, participants were presented with 21 scales that each corresponded to one of 21 values, from 0% to 100% (in 5% intervals; see Figure 2). Each scale had 21-point points and was labeled at each from *not at all* to *absolutely*. Participants were asked: “Decide how well you think each percentage substitutes for the phrase [proof/doubt] doubt that would make a reasonable person [willing/hesitate] to act in their most important affairs of life. In other words, imagine each percentage in front of the word [proof/doubt] and decide to what extent each percentage would make you [willing/hesitate] to act in your most important affairs of life.” Participants responded by circling a point on each scale. As Figure 2 shows, the MF method provides three measures: the ‘minimum’ value that can be substituted for RD, the ‘peak’ value that *absolutely* substitutes for RD, and ‘spread’ of values that represent RD to varying degrees.

FIGURE 2 ABOUT HERE

Procedure

Posters advertising the study were placed across the multi-building site from where the participants were recruited, and emails asking for volunteers were sent out to all employees. Data was collected from individuals in small groups over three days. The experiment took individuals on average 15 minutes to complete.

Participants were asked to imagine that they were serving on a jury in a criminal trial and that after hearing all the evidence for and against the person, the judge instructs them on the standard of proof (there was no case material to consider, however). Each participant read one of the four written RD instructions (i.e., 'proof-willing', 'doubt-hesitate', 'proof-hesitate', and 'doubt-willing'). An equal number of participants read each instruction (i.e., 50 in each condition).

All participants provided their interpretations of RD using the direct rating method. Participants in the 'proof-willing' and 'doubt-hesitate' conditions also provided their interpretations using the MF method, and the order of methods was counter-balanced across these participants. Finally, data was collected on participants' demographic characteristics including their age, gender, educational background, employment status, and jury experience.

Results

Below, we present the analyses and results in order of the four aims listed earlier.

Interpretations of RD under Existing Instructions

The first aim was to compare the effect of the 'proof-willing' and 'doubt-hesitate' instructions on people's quantitative interpretations of RD. Since participants in these two conditions had provided their quantitative interpretations using both the direct rating and MF methods, we first examined the association between the two measures. There was a significant positive correlation of .50 between the point estimate yielded by the direct rating method and the MF peak, $p < .001$. However, a paired-samples *t*-test revealed a significant difference in the mean interpretation provided by the MF peak and the direct rating method, $t[79] = 6.75$, $p < .001$ (see below). Given that the measures provided by the two methods were not fully redundant, we analyzed them separately.

Under the 'proof-willing' instruction, the mean interpretation of RD elicited by the direct rating method was 89.10 ($SD = 8.12$), and the mean (peak) interpretation captured by

the MF method was 94.63 (*SD* = 10.71; see Table 1). Under the 'doubt-hesitate' instruction the mean interpretations of RD according to the direct rating and MF methods were 39.90 (*SD* = 28.04) and 79.13 (*SD* = 24.67), respectively. Independent samples *t*-tests showed that the 'doubt-hesitate' instruction led to significantly lower interpretations of the standard than did the 'proof-willing' instruction both when measured by the direct rating method ($t[98] = 11.92$, $p < .001$) and the MF method, $t(78) = 3.65$, $p < .001$.

TABLE 1 ABOUT HERE

Finally, the MF method also captures the minimum interpretation of RD. As Table 1 shows, the minimum interpretation of RD was significantly lower under the 'doubt-hesitate' than the 'proof-willing' instruction, $t(78) = 4.22$, $p < .001$.

Inter-individual Variability in Interpretations of RD

The second aim was to compare the degree of inter-individual variability in interpretations of RD under the 'proof-willing' and 'doubt-hesitate' instructions. An inspection of the standard deviations of the mean provided by the direct rating method and the mean MF peak in Table 1, shows that there was considerable variability across individuals' interpretations of the standard. A Levene's test revealed that the standard deviation of the mean MF peak was significantly greater under the 'doubt-hesitate' than 'proof-willing' instruction, $F(1, 78) = 19.15$, $p < .001$. Similarly, the standard deviation of the mean interpretation elicited by the direct rating method was significantly greater under the 'doubt-hesitate' than 'proof-willing' instruction, $F(1, 98) = 28.03$, $p < .001$

Intra-individual Variability in Interpretations of RD

Only the MF method provides a measure of intra-individual variability (i.e., spread) of interpretations of RD. As expected, there was considerable intra-individual variability in interpretations of the standard under both the 'proof-willing' and 'doubt-hesitate' instructions (see Table 1). An independent samples *t*-test revealed that the spread of interpretations was

significantly greater under the 'doubt-hesitate' than the 'proof-willing' instruction, $t(78) = 4.19$, $p < .001$.

Effects of 'Willing v. Hesitate to Act' and 'Proof v. Doubt' Language on Interpretations of RD

The final aim was to explain the effect of the two instructions by examining how their language influences interpretations of RD. In particular, we measured how the phrases 'willing to act' and 'hesitate to act' interact with the words 'proof' and 'doubt' to affect interpretations of the standard. An ANOVA was computed with 'willing v. hesitate to act' and 'proof v. doubt' as the two between-subjects factors and interpretations of RD captured by the direct rating method as the dependent measure.

There were significant main effects of both the 'willing v. hesitate to act' phrases and the words 'proof v. doubt' on interpretations of RD, $F(1, 196) = 117.61$, $p < .001$, $\eta_p^2 = .38$ and $F(1, 196) = 6.14$, $p = .014$, $\eta_p^2 = .30$; respectively. Specifically, interpretations were higher under the word 'proof' ($M = 77.35$, $SD = 22.76$) than 'doubt' ($M = 37.30$, $SD = 31.28$). Interpretations of RD were also higher under the phrase 'willing to act' ($M = 61.90$, $SD = 36.91$) than 'hesitate to act' ($M = 52.75$, $SD = 30.03$). However, these main effects were qualified by a significant interaction effect, $F(1, 196) = 15.10$, $p < .001$, $\eta_p^2 = .07$.

The interaction effect is illustrated in Figure 3. From one perspective, this shows that mean interpretations of RD were lower when the phrase 'willing to act' was used in conjunction with the word 'doubt' ($M = 34.70$, $SD = 34.31$) than when 'willing to act' was used in combination with 'proof' ($M = 89.10$, $SD = 8.12$). Similarly, interpretations of RD were lower when the phrase 'hesitate to act' was used in conjunction with the word 'doubt' ($M = 39.90$, $SD = 28.03$) than when it used in combination with 'proof' ($M = 65.60$, $SD = 26.43$). In other words, Figure 3 shows that interpretations of RD were lower when the word

'doubt' was used rather than the word 'proof', regardless of whether 'doubt' was used in conjunction with 'willing to act' or 'hesitate to act' (for doubt and willing to act: M = 34.70, SD = 34.31 for doubt and hesitate to act: M = 39.90, SD = 28.03).

FIGURE 3 ABOUT HERE

Discussion

In the present paper, we go beyond a descriptive account of RD instructions as well as beyond the use of point numerical estimates of RD. We captured the fuzziness of people's interpretations of this standard of proof, and attempted to explain the effect that two RD instructions (i.e., 'proof-willing' and 'doubt-hesitate'), which have not been previously studied, have on interpretations of the standard based on their precise language. As such, our work was informed by the theory of linguistic probabilities (e.g., Budescu & Wallsten, 1995; Dhami, 2008; Wallsten & Budescu, 1995), and built on the psycholinguistic approach to judicial instructions (e.g., Charrow & Charrow, 1979; Elwork et al., 1982; Heffer, 2006; Solan, 2001; Power, 1999).

Like most of the past research on this topic, our evaluation of the effectiveness of RD instructions is premised on the notion that the threshold for conviction should be around .90 (see Blackstone, 1765; Laudan, 2003; McCauliff, 1982; Newman, 1993; United States v. Fatico, 1978). We found a significant difference in the effects of the two RD instructions. The 'doubt-hesitate' instruction lowered the standard considerably, while the 'proof-willing' instruction was interpreted around the desired threshold. This differential effect of the two instructions held both when interpretations were measured using the MF method and the direct rating method.

The above findings provide further evidence for the idea that some RD instructions may not always be effective (see also e.g., Dhami, 2008; Koch & Devine, 1999; Kramer & Koenig, 1990; Montgomery, 1998; Wright & Hall, 2007). In addition, the findings do not

support the judicial preference for the 'doubt-hesitate' instruction (e.g., Holland v. US, 1954; Poulson v. State of Texas, 1999; Wainwright v. US, 1999). Rather, the above findings imply that by reducing the standard of proof below that intended by the law, the 'doubt-hesitate' instruction is more likely to lead to false convictions than the 'proof-willing' instruction. Future research could investigate the effect of the two instructions on both individual (juror) verdict preferences and group (jury) verdicts. Based on the present findings, we would predict that the 'doubt-hesitate' instruction would lead to more convictions.

Perhaps where RD instructions could be of value are in efforts at reducing the inter- and intra-individual variability of interpretations of the standard. However, we found that both RD instructions (i.e., 'proof-willing' and 'doubt-hesitate') induced considerable variability. According to the MF and direct rating methods, the 'doubt-hesitate' instruction led to significantly greater inter-individual variability in interpretations of the standard than did the 'proof-willing' instruction. In addition, according to the MF method, which also captures the spread of an individual's interpretations, the 'doubt-hesitate' instruction led to significantly greater intra-individual variability in interpretations of RD compared to the 'proof-willing' instruction.

The findings on variability of interpretations of RD are consistent with past research on other RD instructions (Dhami, 2008; Mueller-Johnson et al., 2014). This further underscores the fact that RD is a fuzzy concept in jurors' minds (see Clermont, 1987, 2013; Dhami, 2008; Schum, 1986). There are two main (and opposing) implications of these findings. On the one hand, jurors may find it more difficult under the 'doubt-hesitate' than 'proof-willing' instruction to reach consensus on a post-deliberation verdict due to *inter*-individual variability. On the other hand, jurors may be more likely to reach consensus under the 'doubt-hesitate' than 'proof-willing' instruction due to *intra*-individual variability. Future

research could investigate the effect of individual variability in interpretations of RD on group (jury) verdicts.

Finally, we found that the effect of the two RD instructions (i.e., 'proof-willing' and 'doubt-hesitate') could be explained by the precise language used in them. The phrase 'hesitate to act' alone (not in the context of either the words 'proof' or 'doubt') led to lower interpretations of RD than did the phrase 'willing to act'. Similarly, the word 'doubt' (not in the context of either the phrase 'willing to act' or 'hesitate to act') led to lower interpretations of RD compared to the word 'proof'. The significant interaction effect of these phrases and words revealed that when either the phrases 'willing to act' or 'hesitate to act' appear in combination with the word 'doubt' this reduces the standard of proof, compared to when these two phrases appear in conjunction with the word 'proof'. Thus, use of the word 'doubt' explains why the 'doubt-hesitate' instruction led to lower interpretations of RD than the 'proof-willing' instruction.

The implication of our findings is that the use of the word 'doubt' when defining RD ought to be carefully considered, and perhaps even avoided altogether. Indeed, the word 'doubt' is prevalent and focal in many RD instructions. Examples include 'doubt based on reason', 'actual and substantial doubt', and 'doubt that can be articulated' (see Hemmens et al., 1997). Future research ought to examine the effect of the word 'doubt' when used in these other instructions.

One issue that has rarely been considered when discussing the language of RD instructions is the idea that they might (inadvertently) define the juror's (jury's) task. Earlier, we observed that the language of the 'proof-willing' and 'doubt-hesitate' instructions appears to guide jurors in the tasks of evidence evaluation and conviction in different ways. The 'proof-willing' instruction appears to ask jurors to consider evidence as providing proof of guilt and to act to convict. By contrast, the 'doubt-hesitate' instruction appears to ask jurors

to consider evidence as signifying doubt in guilt and not act (or hesitate) to convict. In terms of conviction, the phrases 'willing to act' and 'hesitate to act' appear to ask jurors to perform an action (i.e., to convict) or not to act (i.e., not to convict), respectively. In terms of evidence evaluation, the words 'proof' and 'doubt' appear to ask jurors to consider evidence as providing proof of guilt versus signifying doubt in guilt, respectively. Our findings suggest that the task of evidence evaluation (i.e., as finding 'proof v. doubt') rather than conviction (i.e., 'willing v. hesitate') was particularly important in shaping interpretations of RD.

Despite this, it might be fruitful to consider the task of conviction in future research. Some psychological studies suggest that actions are regretted more than inactions, especially when they lead to bad outcomes (e.g., Kahneman & Tversky, 1982), other research finds a regret inducing effect of inaction, particularly in the long-term (e.g., Gilovich & Medvec, 1994, 1995). Our findings are consistent with the former idea because interpretations of RD were higher under the phrase 'willing to act' than 'hesitate to act', implying that jurors may increase the threshold for conviction (i.e., action) in an effort to reduce any anticipated regret. Future research could examine the effect of anticipated regret (due to an action effect) on interpretations of RD.

One implication is that the task of evidence evaluation is more closely linked to how jurors' think about the standard of proof than the task of conviction itself. Clermont (2013) similarly highlights the importance of the relationship between evidence, beliefs and standards of proof. Future research could explore how definitions of other juror tasks that appear in RD instructions might affect interpretations of the standard. For instance, some RD instructions appear to instruct jurors on how they ought to feel about the defendant and what sort of cognitive approach they ought to apply (e.g., the Supreme Court of Canada in R. Lifchus [1997] states that RD is not based on 'sympathy or prejudice' and that RD should be based on 'reason and commonsense').

Finally, as noted earlier, both the 'proof-willing' and 'doubt-hesitate' instructions are typically used in conjunction with the following language: 'important affairs of life', which we kept constant in the present study. This language is used in around 20 US states, and it has been criticized (see Federal Judicial Centre, 1987; Laudan, 2003). It is believed that decisions people make in their own important affairs are unlike the decision that jurors ought to make in a criminal trial, and that this language trivializes the latter; potentially reducing the standard (e.g., Scurry v. US, 1965). Future research may wish to examine the effect of the 'proof-willing' and 'doubt-hesitate' instructions outside the context of the 'important affairs of life' language.

Further Implications: To Define or Not to Define RD?

Opponents of the idea of defining RD for jurors suggest that leaving the standard undefined avoids the difficulty in attempting to define it and allows the courts to rely on the knowledge of the jury (Notes, 1995). As Heffer (2006, p. 168) points out, judges may advise jurors that the words RD are "ordinary everyday words" and thus should be self-explanatory (see also Power, 1999). Nevertheless, juries sometimes request a definition (Diamond, 1990; Laudan, 2003), and proponents of definition argue that defining the standard of proof for jurors is crucial to a fair trial as it ensures that all jurors understand the criterion for conviction in a similar way (Cohen, 1995). Some commentators have opted for a middle-ground in terms of adding a quantitative definition (Elwork, Alfini & Sales, 1982). However, courts in jurisdictions such as the US and England and Wales have been firmly opposed to quantification (see Laudan, 2003; Power, 1999), and the evidence is mixed as to its effectiveness (Dhami, 2008; Kagehiro, 1990; Kramer & Koenig, 1990).

Although the present findings support the view that qualitative definitions of RD may be problematic, past research has shown that there remains sizeable variability in interpretations of the standard when it is left undefined (e.g., Horowitz & Kirkpatrick, 1996;

Kerr et al., 1976) indicating a lack of understanding, and so need for definition. We argue that it is not definition per se that should be the center of debate, but which language is used to define RD. We propose an evidence-based approach to RD instructions so that their construction is informed by evidence on whether the language in them has desired effects on people's understanding and application of the standard (see also Elwork et al., 1982), and that RD instructions are equally effective for different sub-samples of the jury eligible population (see also Mueller-Johnson et al., 2014).

An evidence-based approach can continue to employ the direct rating and MF methods to measure and validate phrases that can serve as effective definitions of RD. The quantitative interpretations of RD captured by both the direct rating and MF methods have been shown to provide reliable and valid measures of RD as they are associated with theoretically related concepts such as verdicts and verdict preferences (e.g., see Dane, 1985; Dhami, 2008; Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Kerr et al., 1976; Lundrigan et al., 2013). The fact that the two methods led to significantly different mean interpretations of RD in the present study may simply reflect the fact that the direct rating method asks for the *minimum* probability value that represents the standard, whereas the MF method (peak) is the probability value that *absolutely* substitutes for RD (see also Dhami, 2008).

Potential Limitations

It could be argued that the external validity of the present findings is limited because mock jurors were used rather than real jurors; and that we studied the standard of proof outside the context of a legal case, and at the pre-deliberation stage. It would be inappropriate to study real juries in real trial situations where the standard of proof was manipulated experimentally, as we did in the present study. Rather, we used the methodological procedures typical of experimental psychological research on jury decision-making (see

Bornstein & Greene, 2011; Greene, Chopra & Kovera, 2002). Beyond this, we made a concerted effort to minimize the limitations of the method in several ways.

First, most past psychological research on jury decision-making in general and standards of proof in particular, utilizes student samples (for a review see Devine, Clayton, Dunford, Seying & Pryce, 2001). Some studies have shown few differences between mock and real jurors (e.g., MacCoun & Kerr 1988; Reifman, Gusick & Ellsworth, 1992). We sampled participants from a large, multi-building site company. Although this is not equivalent to random sampling from the jury-eligible population (which to our knowledge has never been done in past research on this topic), it did afford the opportunity to study a wide cross-section of people in a controlled data collection environment. In fact, some of the participants had been called for jury service in the past.

Second, while some past research has measured interpretations of RD in the context of simulated criminal cases (e.g., Kagehiro, 1990; Lundrigan et al., 2013), others have not (e.g., Martin & Schum, 1987; Mueller-Johnson et al., 2014). Theoretically speaking, the interpretation of RD should not vary as a function of case, and Dhami (2008) found no significant difference in people's interpretations of RD in and outside the context of a real manslaughter case (see also Lundrigan et al., 2014). If we had studied RD in the context of a criminal case our findings would have been potentially limited to that specific type of case. Future research, nevertheless, could examine the effect of 'proof-willing' and 'doubt-hesitate' instructions in the context of a variety of criminal cases.

Finally, the past research measuring quantitative interpretations of RD has focused on individual-level interpretations (e.g., Koch & Devine, 1999; Wright & Hall, 2007), as we did. There is some evidence that interpretations of RD differ very little from pre- to post-deliberation (Dane, 1985; Horowitz & Kirkpatrick, 1996), and the evidence suggests that juries do not spend a significant amount of time discussing the meaning of RD during

deliberation (e.g., Hastie et al., 1983; Ogloff, 1998). Of course, future research could examine the effect of the 'proof-willing' and 'doubt-hesitate' instructions on deliberating juries.

Conclusion

Given that the legal system confers great responsibility on jurors to make decisions that may have severe consequences for a defendant's liberty and for public security, the system should be responsible for setting out clearly what it asks of the juror, so that jurors can, and are confident that they can, accomplish this task. To date, judicial instructions defining the standard of proof for jurors have not been based on any empirical evidence of their effectiveness. The legal challenges that have resulted from the use of specific definitions have also not been informed by such evidence. We demonstrate the importance of empirical research testing the influence of different definitions of RD, and highlight the powerful effect that linguistic context can have on their efficacy.

References

Blackstone, W. (1765). Commentaries on the Laws of England. Oxford: Clarendon Press.

Bornstein, B. H., & Greene, E. (2011). Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science, 20,* 63–67. doi:10.1177/0963721410397282

Budescu, D. V., Por, H., & Broomell, S. B. (2011). Effective communication of uncertainty in IPCC reports. *Climate Change.* doi: 10.1007/510581-011-0330-3.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, D. L. Medin, & R. Hastie (Eds.), *Decision making from a cognitive perspective* (pp. 275-318). New York: Academic Press.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception & Performance, 14,* 281-294.doi: 10.1037//0096-1523.14.2.281

Charrow, R. E., & Charrow, V. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review, 79*, 1306-1374. doi:10.2307/1121842

Clermont, K. M. (1987). Procedure's magical number three: Psychological bases for standards of decision. *Cornell Law Review, 72,* 1115-1156.

Clermont, K. M. (2013). *Standards of decision in law*. Durham, NC: Carolina Academic Press.

Cohen, J. (1995). The reasonable doubt jury instruction: Giving meaning to a critical concept. *American Journal of Criminal Law, 22*, 677.

Dane, F. C. (1985). In search of reasonable doubt. *Law and Human Behavior, 9*, 141-158. doi: 10.1007/BF01067048

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy and Law, 7*, 622–727. doi:10.1037//1076-8971.7.3.622

Dhami, M. K. (2008). On measuring quantitative interpretations of reasonable doubt. *Journal of Experimental Psychology: Applied, 14*, 353-363. doi:10.1037/a0013344

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: toward translating linguistic probabilities. *Memory and Cognition, 33*, 1057-68. doi: 10.1037/a0013344

Diamond, H. A. (1990). Reasonable doubt: To define or not to define. *Columbia Law Review, 90*, 1716-1736. doi: 10.2307/1122751

Dumas, B. K. (2000). Jury trials: Lay jurors, pattern jury instructions, and comprehension issues. *Tennessee Law Review, 67*, 701-742.

Elwork, A., Alfini, J. J., & Sales, B. D. (1982). Toward understandable jury instructions. *Judicature, 65*, 432-443.

Elwork, A., Sales, B. D., & Alfini, J. J. (1982). *Making jury instructions understandable*. Charlottesville, VA: Michie/Bobb-Merrill.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior & Human Decision Processes, 44*, 1-18. doi: 10.1016/0749-5978(90)90002-Q

Federal Judicial Center (1987). *Pattern criminal jury instructions*. Washington, DC: Federal Judicial Center.

Fortunator, S. J., Jr. (1996). Instructing on reasonable doubt after Victor v. Nebraska: A trial judge's certain thoughts on certainty. *Villanova Law Review, 41*, 365-415.

Gilovich, T., & Medvec, V. H. (1995). The experience of regret: What, when, and why. *Psychological Review, 102*, 379–395. doi: 10.1037/0033-295X.102.2.379

Greene, E., Chopra, S. R., & Kovera, M. B. (2002). Jurors and juries: A review of the field. In J. R. P. Ogloff (Ed.), *Taking psychology and law into the twenty-first century* (pp. 225–284). New York, NY: Kluwer Academic/Plenum.

Harris, A. D. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory & Cognition, 37,* 1571-1578. doi: 10.1037/a0024195

Hastie, R. (Ed.). (1993). *Inside the juror: The psychology of juror decision-making.* Cambridge; Cambridge University Press.

Heffer, C. (2006). Beyond 'reasonable doubt': The criminal standard of proof instructions as communicative act. *The International Journal of Speech, Language and the Law, 13,* 159-188.

Hemmens, C., Scarborough, K. E., & del Carmen, R. V. (1997). Grave doubts about 'reasonable doubt': Confusion in State and Federal courts. *Journal of Criminal Justice, 25,* 231-254. doi: 10.1016/S0047-2352(97)00008-1

Horowitz, I. A. (1997). Reasonable doubt instructions: Commonsense justice and standard of proof. *Psychology, Public Policy, and Law, 3*, 285-302. doi: 10.1037/1076-8971.3.2-3.285

Horowitz, I. A., & Kirkpatrick, L. C. (1996). A concept in search of a definition: the effects of reasonable doubt instructions on certainly of guilt standards and jury verdicts. *Law and Human Behavior, 20*, 655-670. doi: 10.1007/BF01499236

Kagehiro, D. K. (1990). Defining the standard of proof in jury instructions. *Psychological Science, 1*, 194-200. doi: 10.1111/j.1467-9280.1990.tb00197.x

Kagehiro, D. K., & Stanton, W. C. (1985). Legal vs. qualified definitions of standards of proof. *Law and Human Behavior, 9*, 159-178. doi: 10.1007/BF01067049

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American, 246,* 160–173.

Karelitz, T., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied, 10*, 25-41. doi: 10.1037/1076-898X.10.1.25

Kerr, N. L., Atkin, R. S., Stasser, G., Meek, D., Holt, R.W., & Davis, J. H. (1976). Guilt beyond a reasonable doubt: effects of concept definition and assigned decision rule on the judgments of mock jurors. *Journal of Personality and Social Psychology, 43*, 282-394. doi: 10.1037/0022-3514.34.2.282

Koch, C. M., & Devine, D. J. (1999). Effects of reasonable doubt definition and inclusion of a lesser charge on jury verdicts. *Law and Human Behavior, 23*, 653-674. doi: 10.1023/A:1022389305876

Kramer, G. P., & Koenig, D. M. (1990). Do jurors understand criminal jury instructions? Analyzing the results of the Michigan Juror Comprehension Project. *University of Michigan Journal of Law Reform*, 401-437.

Laudan, L. (2003). Is reasonable doubt reasonable? *Legal Theory, 9*, 295-331. doi: 10.1017/S1352325203000132

Lundrigan, S., Dhami, M. K., & Mueller-Johnson, K. (2013). Predicting verdicts using pre-trial attitudes and standard of proof. *Legal and Criminological Psychology*, Online first. doi: 10.1111/lcrp.12043

Lundrigan, S., Dhami, M. K., & Mueller-Johnson, K. (2014). Effect of charge seriousness and consequences of a custodial sentence on juror decision-making. Manuscript submitted for publication.

MacCoun, R. J., & Kerr, N. L. (1988). Asymmetric influence in mock jury deliberation: Jurors' bias for leniency. *Journal of Personality and Social Psychology, 54,* 21-33.

doi: 10.1037//0022-3514.54.1.21

Martin, A. W., & Schum, D. A. (1987). Quantifying burdens of proof: A likelihood ratio approach. *Jurimetrics Journal, 27,* 383–402.

McCauliff, C. M. A. (1982). Burdens of proof: Degrees of belief, quanta of evidence, or constitutional guarantees? *Vanderbilt Law Review, 35*, 1260-1335.

Montgomery, J. W. (1998). The criminal standard of proof. *New Law Journal*, 582-584.

Mueller-Johnson, K., Dhami, M. K., & Lundrigan, S. (2014). Effects of judicial instructions and juror characteristics on interpretations of 'beyond reasonable doubt'. Manuscript submitted for publication.

Nagel, S. S. (1979). Bringing the values of jurors in line with the law. *Judicature, 63,* 189-195.

Newman, J. O. (1993). Beyond "reasonable doubt." *New York University Law Review, 68*, 979-1002. doi: 10.1093/lpr/mgm010.

Notes (1995). Reasonable doubt. *Harvard Law Review, 108*, 1716-1733, 1955-1972.

Ogloff, J. (1998). *Judicial instructions and the jury. A comparison of alternative strategies*. Final report. Vancouver, BC: British Columbia Law Foundation.

Ohio Jury Instructions (2008) The Ohio Judicial Conference. Anderson Publishing.

Pennsylvania Suggested Standard Criminal Jury Instructions (2008) 2nd Ed. The Pennsylvania Bar Institute.

Power, R. C. (1999). Reasonable and other doubts: The problem of jury instructions. *Tennessee Law Review, 67*, 45-123. doi: 10.2139/ssrn.1750267.

Reifman, A., Gusick, S. M., & Ellsworth, P. C. (1992). Real jurors' understanding of the law in real cases. *Law and Human Behavior, 16*, 539-554. doi: 10.1007/BF01044622.

Schum, D. (1986). Probability and the processes of discovery, proof, and choice. *Boston University Law Review, 66*, 825-876.

Smithson, M., Budescu, D. V., Broomell, S. B., & Por, H. (2012). Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning*. doi: 10.1016/j.ijar.2012.06.019.

Solan, L. M. (2001). Refocusing the burden of proof in criminal cases. *Texas Law Review, 78,* 105-147.

Stoffelmayr, E., & Diamond, S. S. (2000). The conflict between precision and flexibility in explaining" beyond a reasonable doubt". *Psychology, Public Policy, and Law*, *6*(3), 769.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review, 10,* 43-62. doi: 10.1017/S0269888900007256

Wallsten, T. S., Fillenbaum, S., & Cox, A. (1986). Base-rate effects on the interpretations of probability and frequency expressions. *Journal of Memory & Language, 25,* 571-587. doi: 10.1016/0749-596X(86)90012-4

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception & Performance, 16,* 781-789. doi: 10.1037//0096-1523.16.4.781

Wright, D. B., & Hall, M. (2007). How a "reasonable doubt" instructions affects decision of guilt. *Basic and Applied Social Psychology, 29*, 91-98. doi: 10.1080/01973530701331254.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353. doi: 10.1016/S0019-9958(65)90241-X

Zander, M. (2000). The criminal standard of proof – how sure is sure? *New Law Journal, 150,* 1517-1519.

Zwick, R., & Wallsten, T. S. (1989). Combining stochastic uncertainty and linguistic inexactness: theory and experimental evaluation of four fuzzy probability models. *International Journal of Man-Machine Studies, 30,* 69–111. doi: 10.1016/S0020-7373(89)80021-5

Cases Cited

Hopt v. State of Utah, 1887

Posey v. State of Alabama, 1922

Holland v. US 1954

Scurry v. US 1965

United States v. Emalfarb, 1973

United States v. Mars, 1977

United States v. Fatico 1978

Cage v. Louisiana, 1990

Sullivan v. Louisiana, 1993

Gundy v. Ohio 1992

Victor v. Nebraska, 1994

United States v. Miller, 1996

R v. Lifchus 1997

Poulson v. Texas 1999

Wainwright v. US 1999

Harris v. Idaho 2001

Laramore v. Idaho 2007

Cooper v. Alaska 2008

Table 1. Means and Standard Deviations of Interpretations of RD under Existing Judicial Instructions

	M (*SD*)	
	Proof-Willing (*n* = 50)	Doubt-Hesitate (*n* = 50)
Direct Rating	89.10 (8.12)	39.90 (28.04)
MF Peak	94.63 (10.71)	79.13 (24.67)
MF Minimum	41.38 (25.97)	19.38 (20.36)
MF Spread	58.50 (26.02)	80.38 (20.30)

Note. Ten participants did not provide complete data for the MF method.

Figure Captions

Figure 1. Example membership function for RD

Figure 2. Membership function method

Figure 3. Mean interpretations of RD under the ‘willing-hesitate to act’ phrases and ‘proof-doubt’ words

Figure 1

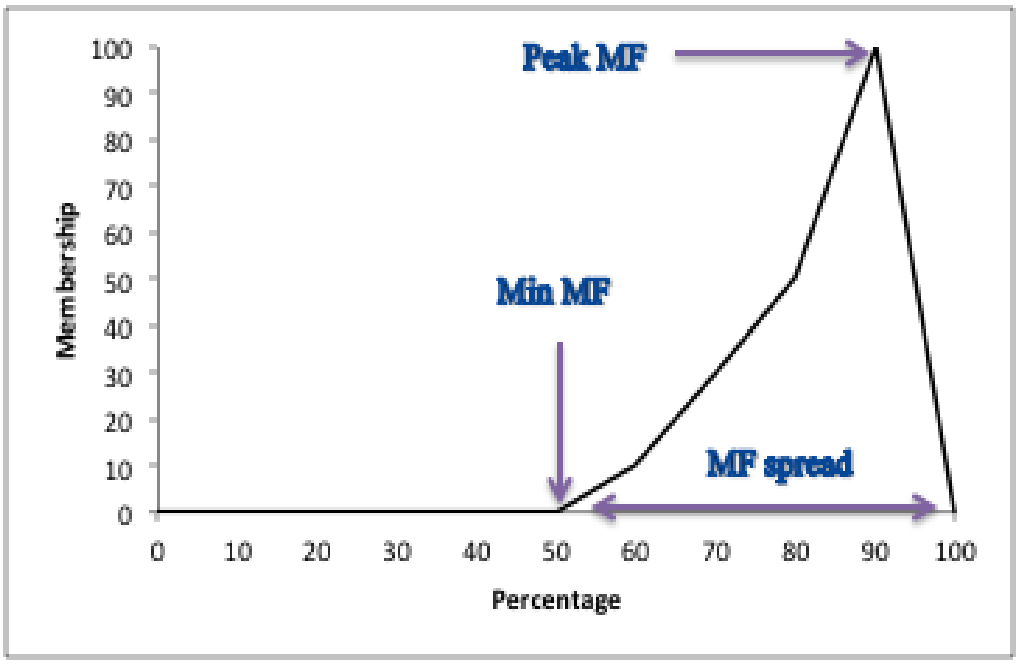

Peak MF
Min MF
MF spread
Membership
Percentage
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100

Figure 2

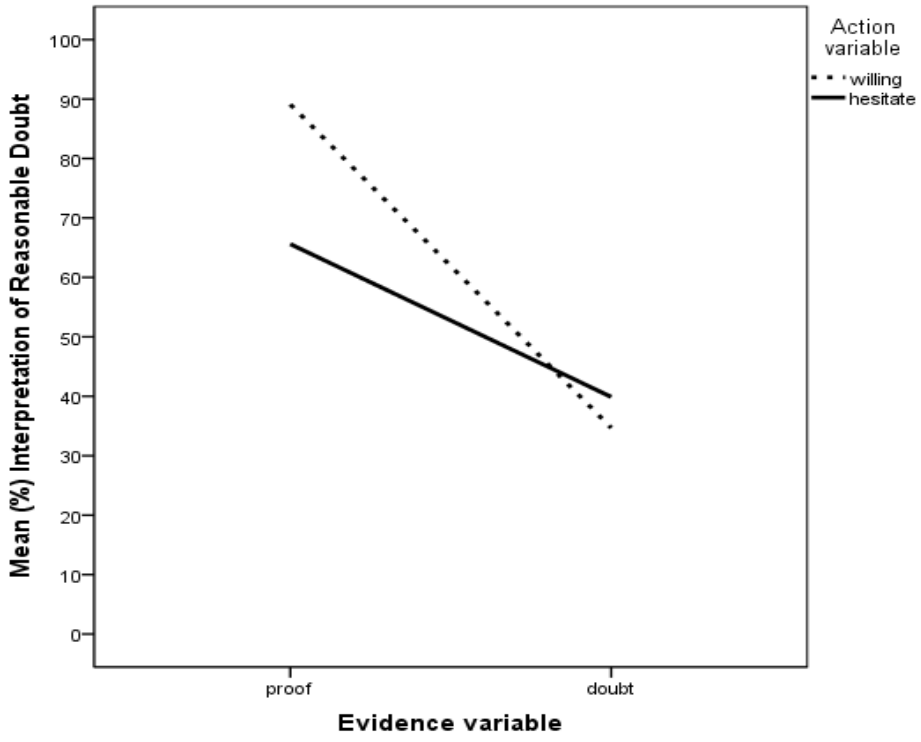
Statement: "The defendant is presumed innocent unless the prosecution has proved guilt beyond reasonable doubt. Reasonable doubt is doubt that would make a reasonable person hesitate to act in their most important affairs of life."

Below is a list of percentages. Next to each percentage is a scale from *not at all* to *absolutely*. Decide how well you think each percentage substitutes for the phrase doubt that would make a reasonable person hesitate to act in their most important affairs of life. In other words, imagine each percentage in front of the word 'doubt' and decide to what extent each percentage would make you hesitate to act in your most important affairs of life.

For example, if you think that 0%, 5% and 10% would *not at all* make you hesitate then circle the left-most point on those scales. If you think 90% would *absolutely* make you hesitate then circle the right-most point on the scale. And, if you think 80% would make you hesitate more than 70%, then circle a point along the 80% scale that is closer to absolutely than the point you circle along the 70% scale. Make sure you circle one point on each scale.

0%	Not at all	Absolutely
5%	Not at all	Absolutely
10%	Not at all	Absolutely
15%	Not at all	Absolutely
20%	Not at all	Absolutely
25%	Not at all	Absolutely
30%	Not at all	Absolutely
35%	Not at all	Absolutely
40%	Not at all	Absolutely
45%	Not at all	Absolutely
50%	Not at all	Absolutely
55%	Not at all	Absolutely
60%	Not at all	Absolutely
65%	Not at all	Absolutely
70%	Not at all	Absolutely
75%	Not at all	Absolutely
80%	Not at all	Absolutely
85%	Not at all	Absolutely
90%	Not at all	Absolutely
95%	Not at all	Absolutely
100%	Not at all	Absolutely

Figure 3


Mean (%) Interpretation of Reasonable Doubt
100
90
80
70
60
50
40
30
20
10
0
proof
doubt
Evidence variable
Action variable
willing
hesitate